Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Information Theoretical Aspects of Complex Systems

Lecture 2.05

EEU45C09 / EEP55C09

Self Organising Technological Networks

CONNECT Networks of the Future

Nicola Marchetti
nicola.marchetti@tcd.ie

Entropy vs redundancy

- Information theory was developed to deal with signals or sequences of symbols

- Information quantity can be decomposed in two terms:

  - Entropy term that quantifies the system's *disorder*
  - Information that quantifies the system's *order* (usually called *redundancy*)

- Entropy quantifies the uncertainty that remains (on average), when all correlations have been taken into account before observing the next symbol

- This is lack of knowledge, meaning it is in the entropy part of the sequence where we can transmit information between Transmitter and Receiver

Entropy vs redundancy

- Based on statistics from a source text, we form conditional probabilities $p(x_n|x_1,\cdots,x_{n-1})$, expressing the probability for the next symbol given the n-1 preceding ones

- Using these probabilities, we randomly generate symbols, one by one

- The longer the correlation we decide to include (larger n), the longer the preceding sequence we take into account when calculating the probability

- In the following six examples we go from a text generated using only n = 1 (we call this *density information*) to texts with correlations over block lengths 2-6

Entropy vs redundancy

1: Tdory d neAeeeko,hs wieadad ittid eIa c i lodhign un a a svmb i ee' kwrdmn.

2: Le hoin. whan theoaromies out thengachilathedrid be we frergied ate k y wee ' e the sle! se at te thenegeplid whe tly titou hinyougea g l fo nd

3: 'Weed. Thed to dre you and a dennie. A le men eark yous, the sle nown ithe haved saindy. If - it to to it dre to gre. I wall much. 'Give th pal yould the it going, youldn't thave away, justove mouble so goink steace, 'If take we're do mennie.

4: I can light,' George tried in you and fire.' 'Nothen it and I want yourse, George some other ther. There's if his hand rolledad ther hisky, 'I little amonely we're we're with him the rain.

5: 'I...I'm not running.' The ranch, work on the time. Do you because you get somethings spready told you just him by heat to coloured rabbits. That's going grew it's like a whisky, place.

6: Million mice because it two men we'll sit by the future. We'll steal it. 'Aren't got it. 'About the fire slowly hand. 'I want, George,' he asked nervoulsly: 'That's fine. Say it too hard forget other.

Entropy vs redundancy

- Already with correlation over n = 2 or n = 3 , it is clear this is English
- When correlation (n) is increased, more and more words are correctly generated
- With n = 6 , we might even guess what story the text is taken from
- We will now present a *formalism that can be used to analyse disorder and correlations* in symbol sequences
- We will first start with 1 dimensional *lattices*, then extend to 2 dimensions
- We will then use the formalism to analyse states (i.e. symbol sequences) in the time evolution of *discrete dynamical systems (cellular automata)*

One-dimensional lattices

- Let us assume the system under study consists of infinite sequences of symbols, where each symbol is taken from some finite alphabet Λ

- The system is a set (ensemble) of such sequences

- We may form probability distributions over sets of finite sub-sequences of symbols, one for each length n (n = 1, 2, 3, …)

- We consider symbol sequences being generated by a *stationary* stochastic process, which can be described by an infinite sequence of stochastic variables X_k

- Because of stationarity, probabilities of sub-sequences do not depend on position, but only on how many former observations I did (here X_n is the outcome of a certain experiment at time n):

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{1+m} = x_1, X_{2+m} = x_2, \ldots, X_{n+m} = x_n)$$

for all n, m, and any symbols $x_k \in \Lambda$.

One-dimensional lattices

- We can characterise our system by a probability distribution P_n over symbol sequences of finite length n

$$P_n = \left\{ p_n(x_1, \ldots, x_n) \right\}_{x_1, \ldots, x_n \in \Lambda^n} \qquad (n = 1, 2, \ldots)$$

$$P_n = \left\{ p_n(\sigma_n) \right\}_{\sigma_n \in \Lambda^n} \qquad (n = 1, 2, \ldots)$$

- There are conditions that relate probability distributions over lengths n and n + 1 , based on the fact that the distribution over (n+1)-length sequences includes the distribution over n-length sequences

$$p_n(x_1, \ldots, x_n) = \sum_{x_{n+1} \in \Lambda} p_{n+1}(x_1, \ldots, x_n, x_{n+1}) \text{ , and}$$

$$p_n(x_1, \ldots, x_n) = \sum_{x_0 \in \Lambda} p_{n+1}(x_0, x_1, \ldots, x_n) \text{ .}$$

Shannon entropy

- *A priori knowledge* of the system
  - Each symbol belongs to a certain alphabet Λ
  - Λ contains $|\Lambda| = \nu$ different characters
  - Our initial uncertainty (*lack of knowledge*) per symbol is then $S = \log \nu$

- Successively adding probability distributions for sequences of increasing length P_n $(n = 1,2,\ldots)$, we take *correlation* into account to reduce uncertainty of the next symbol in the sequence

- The entropy that may still remain when we include correlations over all lengths ($n \to \infty$) is the *Shannon entropy*

Mutual entropy

- Before we formalise this discussion, let us consider the following example involve two, possibly dependent, stochastic variables X_1 and X_2 , where their possible outcomes are characters belong to the alphabet Λ

- If we first assume they are independent and equally distributed, then

$$S[X_1, X_2] = S[X_1] + S[X_2]$$

- If there is a ***correlation (dependence)*** between X_1 and X_2 , then their combined (mutual) entropy $S[X_1, X_2]$ should be less than $S[X_1] + S[X_2]$

Mutual information

See Lecture 2.03

$$I[X_1;X_2] = S[X_1] + S[X_2] - S[X_1,X_2] =$$

$$= \sum_{x_1} p(x_1)\log\frac{1}{p(x_1)} + \sum_{x_2} p(x_2)\log\frac{1}{p(x_2)} - \sum_{x_1x_2} p(x_1x_2)\log\frac{1}{p(x_1x_2)} =$$

$$= \sum_{x_1x_2} p(x_1x_2)\log\frac{p(x_1x_2)}{p(x_1)p(x_2)} = K[P(X_1)P(X_2);P(X_1X_2)] \geq 0$$

Prove this

- The *mutual information* is the information we get when we replace the separate distributions $P(X_1)$ and $P(X_2)$ as the description of the system, with the correct joint distribution $P(X_1,X_2)$

Mutual information

$$I[X_1;X_2] = \sum_{x_1 x_2} p(x_1)\frac{p(x_1x_2)}{p(x_1)}\log\frac{p(x_1x_2)}{p(x_1)p(x_2)} =$$

$$= \sum_{x_1 x_2} p(x_1)p(x_2 \mid x_1)\log\frac{p(x_2 \mid x_1)}{p(x_2)} =$$

$$= \sum_{x_1} p(x_1)K[P(X_2);P(X_2 \mid x_1)].$$

- The Kullback information quantifies the information gained when we replace the probability distribution $P(X_2)$ with the conditional probability distribution that includes the possible correlation that may exist when we have already observed a specific outcome x_1 of X_1

- Then the average over the possible outcomes of X_1 is calculated

- This is also an intuitively reasonable interpretation of the mutual information quantity

Block entropy

- The analysis of correlations in symbol sequences builds on how the entropy of n-length distributions varies with length n

- *Block entropy*

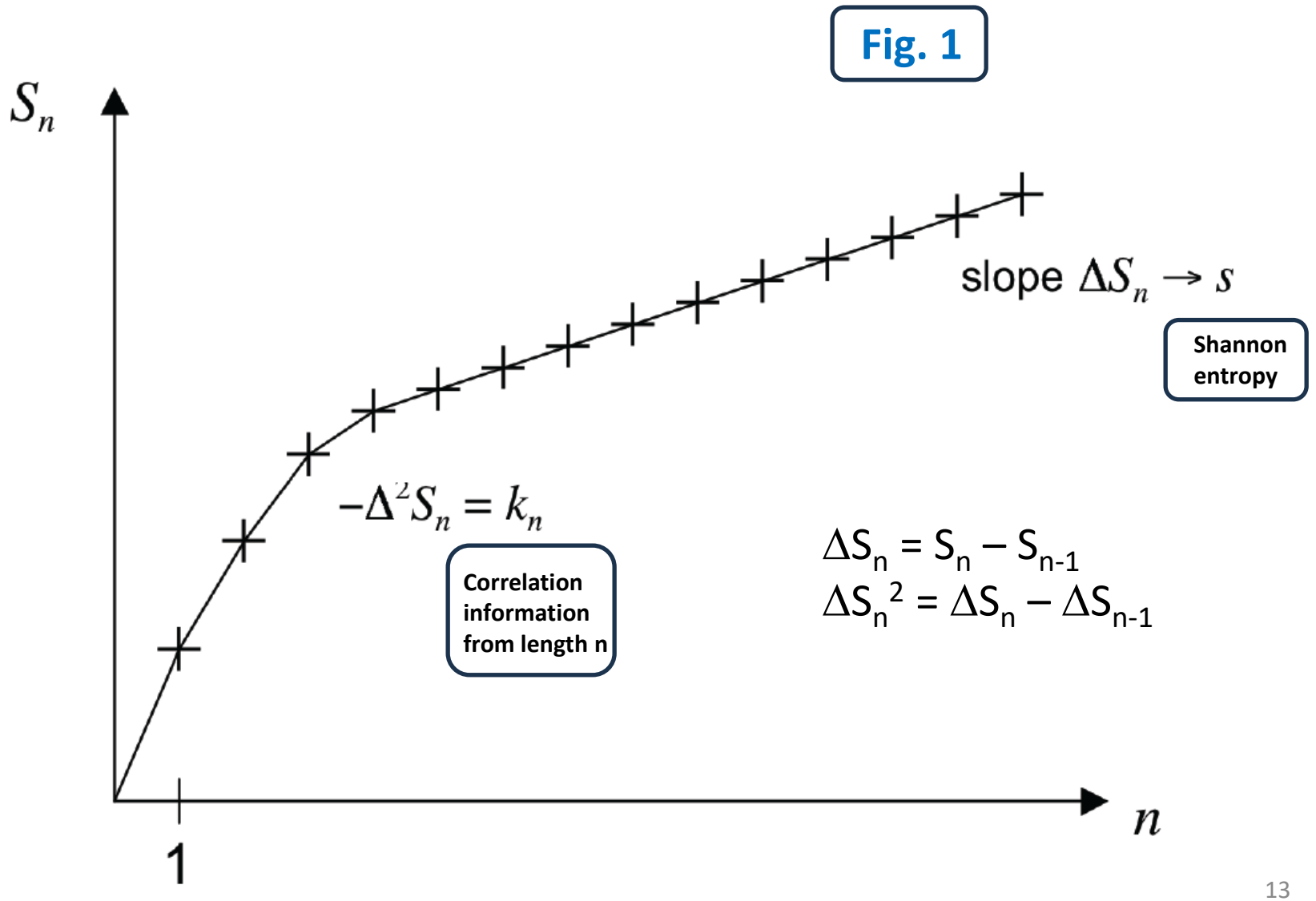
$$S_n = S[P_n] = \sum_{\sigma_n} p(\sigma_n) \log \frac{1}{p(\sigma_n)}$$

- Quantifies disorder of n-length sub-sequences of the system

- S_n grows with n , but if there are correlations then the increase is less than the entropy of an isolated symbol

- The larger n , the longer correlations can be taken onto account, and thus the increase of S_n will be smaller in that case

Block entropy



Correlation information

- For each possible preceding sequence $(x_1, \cdots, x_{n-1})$ we can write the conditional probability of a symbol x_n , given that we have already observed those n-1 preceding symbols

$$p(x_n \mid x_1 \ldots x_{n-1}) = \frac{p(x_1 \ldots x_{n-1} x_n)}{p(x_1 \ldots x_{n-1})}$$

- The entropy for this distribution is a measure of the difficulty in guessing the next symbol, based on the fact that we already know the preceding n-1 ones

$$\left\langle S[P(\bullet \mid x_1 \ldots x_{n-1})\right\rangle = \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum_{x_n} p(x_n \mid x_1 \ldots x_{n-1}) \log \frac{1}{p(x_n \mid x_1 \ldots x_{n-1})} =$$

$$= S_n - S_{n-1} = \Delta S_n \ .$$

Prove this

- This shows that the average conditional entropy equals the slope of the block entropy S_n as a function of length n (see Fig. 1)

Correlation information

- If we increase the length of the preceding symbol sequence in the conditional probability distribution, we may increase our chances to make a better estimate of the probability for the next symbol

- This is due to the fact that increasing the length of the block includes more correlations in the system, and these correlations can be used when guessing the next symbol

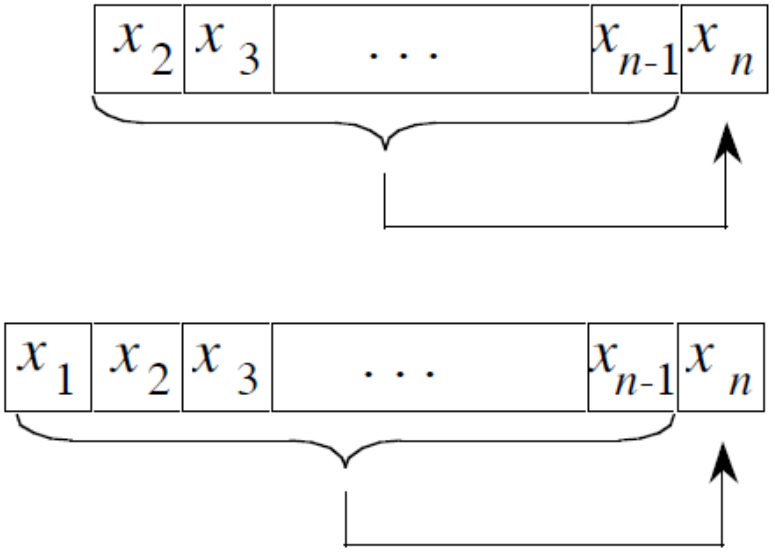


Fig. 2

Correlation information

- To quantify the information in correlations of length n, suppose that we have an a priori conditional distribution $P^{(0)}(\bullet|x_2,\ldots,x_{n-1})$ for a symbol x_n, given that a specific preceding sequence $(x_2,\ldots,x_{n-1})$ is known

- Then we are interested in the information we get when we observe the symbol x_1 and use that to change our probability description of the symbol x_n to $P(\bullet|x_1,x_2,\ldots,x_{n-1})$, see Fig. 2

- The conditional probabilities in $P^{(0)}$ cannot include any correlations of length n (stretching over a sequence of length n), but that is possible in the new distribution P

Correlation information

- The Kullback information between $P^{(0)}$ and P is then a measure of the correlation information of length n when a specific preceding sequence $(x_1, x_2, \ldots, x_{n-1})$ is observed

$$K[P^{(0)};P] = \sum_{x_n} p(x_n \mid x_1 x_2 \ldots x_{n-1}) \log \frac{p(x_n \mid x_1 x_2 \ldots x_{n-1})}{p(x_n \mid x_2 \ldots x_{n-1})}$$

- If we now take the average over all possible preceding sequences $(x_1, x_2, \ldots, x_{n-1})$, we get an expression for the average information content k_n in correlations of length n

- This quantity can be rewritten in the form of a Kullback information, the *correlation information from length n*

Correlation information

- *Correlation information from length n*

$$k_n = \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) K[P^{(0)}; P] =$$

$$= \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum_{x_n} \frac{p(x_1 \ldots x_{n-1} x_n)}{p(x_1 \ldots x_{n-1})} \log \frac{p(x_1 \ldots x_{n-1} x_n) p(x_2 \ldots x_{n-1})}{p(x_1 \ldots x_{n-1}) p(x_2 \ldots x_{n-1} x_n)}$$

$$k_n = -S_n + 2S_{n-1} - S_{n-2} = -\Delta S_n + \Delta S_{n-1} = -\Delta^2 S_n \quad (n = 2, 3, \ldots)$$

See Fig. 1

Prove this

- Correlation information from length 1

$$k_1 = K[P_1^{(0)}; P_1] = \sum_{x_1} p(x_1) \log \frac{p(x_1)}{1/\nu} = \log \nu - S_1$$

Correlation information

- k_1 is an information quantity that measures the difference in character frequency from a uniform distribution

- As an a priori distribution we use the completely "uninformed" uniform distribution $P_1^{(0)}$ that assigns equal probabilities $p_1^{(0)} = 1/\nu$, to all characters x_1 in Λ

- The density information k_1 can then be written as a Kullback information between the a priori uniform distribution and the observed single character distribution P_1

- We have thus defined a number of information quantities that captures the ordered information in the system — the density information k_1 and the series of correlation information contributions k_n (n = 2, 3, …). Let us combine all these into an information quantity, the correlation information k_{corr}

Correlation information

- *Correlation information*

$$k_{\text{corr}} = \sum_{m=1}^{\infty} k_m = \log \nu - \lim_{m \to \infty} (S_{m+1} - S_m) = \log \nu - \Delta S_{\infty}$$

- This is the ordered part (*redundancy*) of the information in the system (expressed as an average per symbol)

Shannon entropy

- The entropy per symbol of the system – the *Shannon entropy s* – is a measure of the uncertainty that remains when all correlations have been taken into account

- One way to define this quantity - the remaining uncertainty - is to use the entropy of the conditional probability for the next character, given that we have already observed a preceding sequence of symbols

- Then we take the average of this entropy over the possible preceding sequences, and we take the infinite limit of the length of the preceding sequence to include all possible correlations in the conditional entropy

- $$s = \lim_{n\to\infty} \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum_{x_n} p(x_n \mid x_1 \ldots x_{n-1}) \log \frac{1}{p(x_n \mid x_1 \ldots x_{n-1})} = \lim_{n\to\infty} \Delta S_n = \Delta S_\infty$$

Maximum entropy = redundancy + disorder

- We can conclude that the total entropy per symbol ($\log \nu$) – which is the maximum entropy per symbol – can be decomposed in two terms, the redundancy k_{corr} and the Shannon entropy s

$$S_{\max} = \log \nu = (\log \nu - \Delta S_{\infty}) + \Delta S_{\infty} = k_{\text{corr}} + s$$

Redundancy

Shannon entropy

Ackowledgement

- Kristian Lindgren, "Information Theory for Complex Systems", pages 16-24